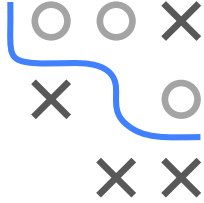
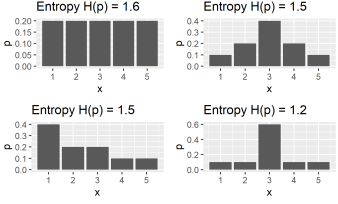
Introduction to Machine Learning



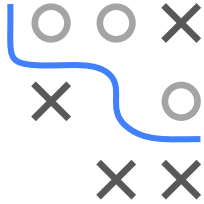
Entropy I



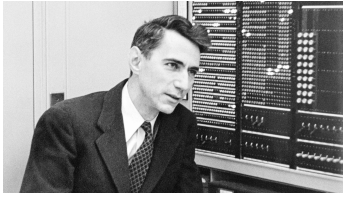
Learning goals

- Entropy measures expected information for discrete RVs
- Know entropy and its properties

INFORMATION THEORY

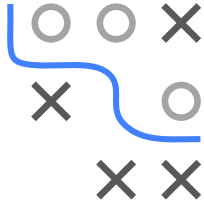


- **Information Theory** is a field of study based on probability theory.
- The foundation of the field was laid by Claude Shannon in 1948 and it has since found applications in areas as diverse as communication theory, computer science, optimization, cryptography, machine learning and statistical inference.
- In addition to quantifying information, it also deals with efficiently storing and transmitting the information.
- Information theory tries to quantify the "amount" of information gained or uncertainty reduced when a random variable is observed.

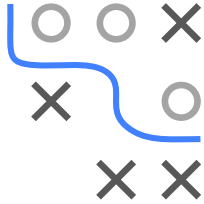


INFORMATION THEORY

- We introduce the basic concepts from a probabilistic perspective, without referring too much to communication, channels or coding.
- We will show some proofs, but not for everything. We recommend *Elements of Information Theory* by Cover and Thomas as a reference for more.
- The application of information theory to the concepts of statistics and ML can sometimes be confusing, we will try to make the connection as clear as possible.



ENTROPY



- We develop in this unit entropy as a measure of uncertainty in terms of expected information.

For a discrete random variable X with domain $\mathcal{X} \ni x$ and pmf $p(x)$:

$$H(X) := H(p) = -\mathbb{E}[\log_2(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

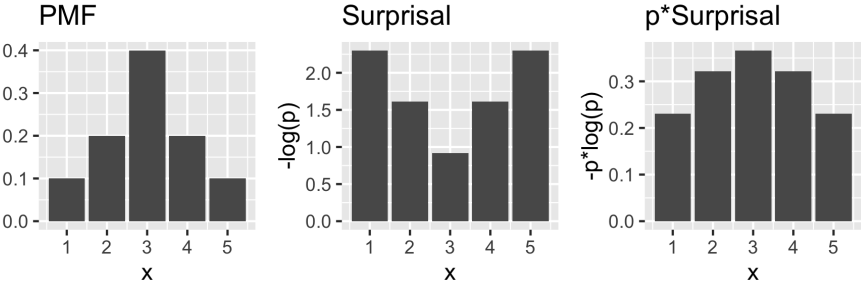
$$= \mathbb{E}\left[\log_2\left(\frac{1}{p(X)}\right)\right] = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)}$$

- **Definition:** Base 2 means the information is measured in bits, but you can use any number > 1 as base of the logarithm.
- **Note:** If $p(x) = 0$, then $p(x) \log_2 p(x)$ is taken to be zero, because $\lim_{p \to 0} p \log_2 p = 0$.
- NB: H is actually Greek capital letter **E**ta (η) for **e**ntropy

ENTROPY CALCULATION

- The negative log probabilities $\log_2 p(x)$ are called "Surprisal".

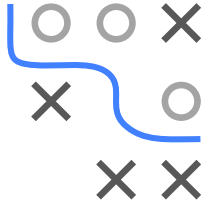
$$H(X) = -\mathbb{E}[\log_2(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$



- The final entropy is $H(X) = 1.5$.

ENTROPY PROPERTIES

$$H(X) := H(p) = -\mathbb{E}[\log_2(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$



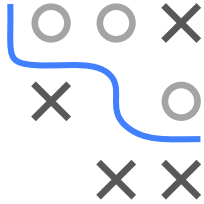
We can directly note some basic properties:

1. Entropy is non-negative, so $H(X) \geq 0$.
2. If one event has probability $p(x) = 1$, then $H(X) = 0$.
3. Adding or removing an event with $p(x) = 0$ does not change entropy.
4. $H(X)$ is continuous in probabilities $p(x)$.

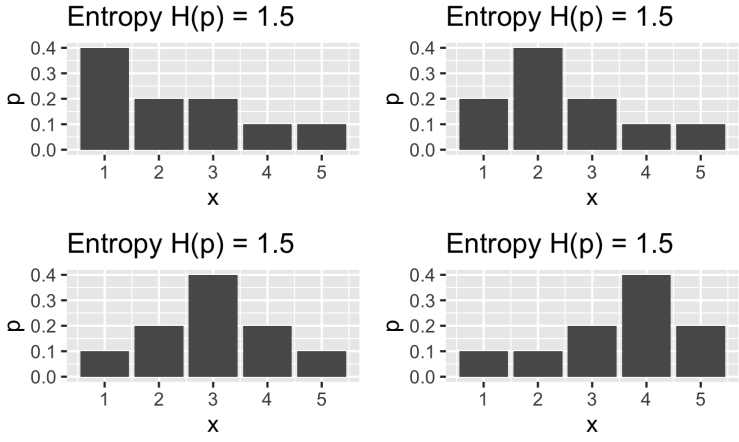
All these properties follow directly from the definition.

In the following, we will look at various simple examples are derive some more properties of the entropy.

ENTROPY RE-ORDERING



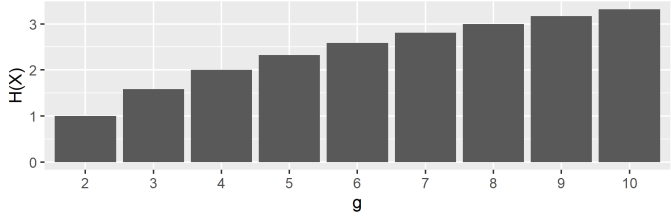
5. Symmetry. If the values $p(x)$ in the pmf are re-ordered, entropy does not change (proof is trivial).



ENTROPY OF UNIFORM DISTRIBUTIONS

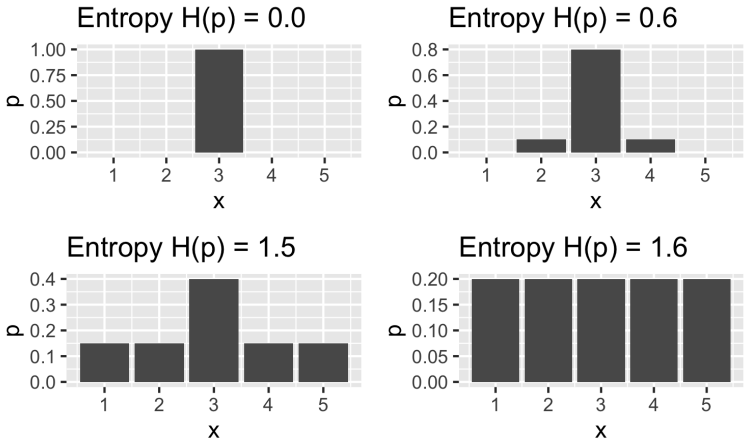
Let X be a uniform, discrete RV with g outcomes (g-sided fair die).

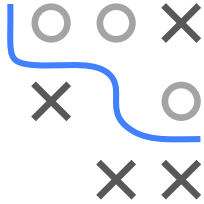
$$H(X) = -\sum_{i=1}^{g} \frac{1}{g} \log_2 \left(\frac{1}{g} \right) = \log_2 g$$



The more sides a die has, the harder it is to predict the outcome. Unpredictability grows *monotonically* with the number of potential outcomes, but at a decreasing rate.

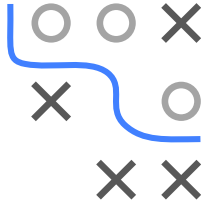
ENTROPY IS MAXIMAL FOR UNIFORM





- Naive observation: Entropy minimal for peaked distribution and maximal for uniform distribution.

ENTROPY IS MAXIMAL FOR UNIFORM



6. Entropy is maximal for a uniform distribution, so for a domain with g elements: $H(X) \leq -g\frac{1}{g}\log_2(\frac{1}{g}) = log_2(g)$.

Claim: The entropy of a discrete random variable X which takes on values in $\{x_1, x_2, \ldots, x_g\}$ with associated probabilities $\{p_1, p_2, \ldots, p_g\}$ is maximal when the distribution over X is uniform.

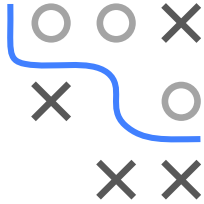
Proof: The entropy $H(X)$ is $-\sum_{i=1}^{g} p_i \log_2 p_i$ and our goal is to find:

$$\underset{p_1,p_2,\ldots,p_g}{\text{argmax}} \; -\sum_{i=1}^{g} p_i \log_2 p_i$$

subject to

$$\sum_{i=1}^{g} p_i = 1.$$

ENTROPY IS MAXIMAL FOR UNIFORM



The Lagrangian $L(p_1, \ldots, p_g, \lambda)$ is :

$$L(p_1, \ldots, p_g, \lambda) = -\sum_{i=1}^{g} p_i \log_2(p_i) - \lambda \left(\sum_{i=1}^{g} p_i - 1 \right)$$

Solving for $\nabla L = 0$,

$$\frac{\partial L(p_1, \ldots, p_g, \lambda)}{\partial p_i} = 0 = -\log_2(p_i) - 1 - \lambda$$
$$\implies p_i = 2^{(-1-\lambda)} \implies p_i = \frac{1}{g},$$

where the last step follows from the fact that all p_i are equal and the constraint.

NB: We also could have solved the constraint for p_1 and substitute $p_1 = 1 - \sum_{i=2}^{g} p_i$ in the objective to avoid constrained optimization.